APPROACHES TO ARCHIVING PROFESSIONAL BLOGS HOSTED IN THE CLOUD

Brian Kelly and Marieke Guy
UKOLN, University of Bath,
Claverton Road, Bath, UK

ABSTRACT

Early adopters of blogs will have made use of externally-hosted blog platforms, such as Wordpress.com and Blogger.com, due, perhaps, to the lack of a blogging infrastructure within the institution or concerns regarding restrictive terms and conditions covering use of such services. There will be cases in which such blogs are now well-established and contain useful information not only for current readership but also as a resource which may be valuable for future generations.

The need to preserve content which is held on such third-party services ("the Cloud') provides a set of new challenges which are likely to be distinct from the management of content hosted within the institution, for which institutional policies should address issues such as ownership and scope of content. Such challenges include technical issues, such as the approaches used to gather the content and the formats to be used and policy issues related to ownership, scope and legal issues.

This paper describes the approaches taken in UKOLN, an applied research department based at the University of Bath, to the preservation of blogs used in the organisation. The paper covers the technical approaches and policy issues associated with the curation of blogs a number of different types of blogs: blogs used by members of staff in the department; blogs used to support project activities and blogs used to support events.

1. BLOG USAGE WITHIN UKOLN

UKOLN is a national centre of expertise in networked information management based at the University of Bath. Our interest in innovation may require staff to use services, such as blogs, which are not provided within our organisation or by our host institution.

Since UKOLN has interests in digital preservation we seek to ensure that we use our experiences to inform best practices on long term access to content held on such services. Such experiences are beneficial in our role in advising UK higher educational institutions on best practices related to use of new and emerging technologies. This paper describes the approaches we have taken and provides advice for other institutions which may have similar concerns.

2. CASE STUDIES

This paper describes three scenarios illustrating differing uses of blogs in UKOLN and highlights the challenges the examples provide regarding the preservation of the contents of blogs.

2.1. The Professional's Blog

The UK Web Focus blog (see http://ukwebfocus.wordpress.com/) was established by Brian Kelly in November 2006. Although there had been some previous experimentation with use of blogs this was the first high-profile blog to be provided by a member of staff and endorsed by JISC (UKOLN's core funding organisation) as a key user engagement and dissemination channel for aspects of UKOLN's work. Since at the time the blog was established neither UKOLN nor the University of Bath provided a blogging platform the WordPress.com service was selected to host the blog.

Since its launch over 750 posts have been published (an average of about four per week) and the blog has attracted over 250,000 user visits (an average of about 240 per day).

This blog supports the author's professional activities and is also written in a personal style which reflects the author's interests and personality. The same is true of Marieke Guy's Rambling of a Remote Worker blog (see http://remoteworker.wordpress.com/).

These two examples illustrate how there may be a degree of uncertainty as to whether the blog posts should be regarded as having institutional or personal ownership.

In light of the popularity and significance of the blog it has been recognised that there is a need to ensure that best practices are developed in order to minimize the risks associated with use of a third-party service to host the content and the risks of loss of institutional IPR which is managed by the blog author, with no formal mechanism for access by others in the institution and no well-understood levels of accountability for the curation of the content by the author.

In addition to clarity regarding such responsibilities there is a need to identify the tools and processes for curating the blog's content independently from the existing platform and, possibly, ownership.

2.2. The Project Blog

The JISC PoWR (Preservation of Web Resources) project was funded by the JISC and provided by a partnership of UKOLN and ULCC. The project ran from April – November 2008. A WordPress blog was used to support the project work which was hosted by the JISC on their JISC Involve platform (see http://jiscpowr.jiscinvolve.org/).

Content for the blog was provided by staff from the two partner organisations. In order to avoid possible confusions regarding ownership of the content it was agreed that blog posts would be published under a Creative Commons licence and a statement to this effect was provided on the blog.

A decision was made to host the blog on a platform provided by the project's funding body rather than using the host institution of either of the project partners. Although this should avoid the risk of unanticipated changes to terms and conditions for the service we are aware that expected cuts in funding for higher education could result in withdrawal of the service or a failure for the service to be developed. We therefore have an interest in the migration of the content of the blog in the unlikely situation that such changes do occur.

2.3. The Event Blog

UKOLN's Institutional Web Management Workshop (IWMW) is an annual 3-day event. The event provides an opportunity to demonstrate uses of innovative Web technologies. After use of wikis and social networking services in previous years in 2009 the choice was made to use an externally-hosted WordPress.com blog (see http://iwmw2009.wordpress.com/).

In addition to posts from the organisers, speakers and other participants at the event were invited to contribute to the blog. Interviews with participants were also published on the blog both as text and video interviews.

As well as embedded video clips (which are hosted on the Vimeo video sharing service) the blog also provided embedded photographs taken at the event which are hosted on Flickr.

This event blog has given rise to some additional challenges related to the long-term preservation of the content including the ownership of content provided by contributors who do not work for UKOLN, privacy issues related to hosting photographs of participants at the event and the sustainability of the content hosted on other third party services.

3. WHAT ARE THE REQUIREMENTS?

Although three different use cases for organisational blogs have been provided it is not necessarily the case that the same requirements will be needed for the 'archiving' of the blogs. It should be noted that the 'archiving' term is being used to describe ways in which blog content can be migrated to alternative environments in order to satisfy a number of business functions, including the re-creation of the original environment. A number of approaches have been identified which are relevant to our use cases:

Production of a new static master version of the content: This approach is felt to be appropriate for use of project blogs when the project has ceased. The contents of the blog can be migrated as static HTML pages. In order to avoid confusion with multiple copies of the content being available the original blog may have a pointer to the new static resource, possibly with the original content being removed from public view.

Production of a backup version of the content: There may be a need to ensure a backup copy of a blog is available in order to avoid the risks of loss of data if the hosting service is not sustainable or, if as has been seen in the case of the Theoretical Librarian blog (which was hosted at http://theoretical-librarian.blogspot.com/) a blog is removed by the service provider, as illustrated in Figure 1.

Figure 1: Removal of a Blog at Blogger.com

Migration of the rich content to an alternative platform: It may be felt necessary to migrate the contents of a blog to an alternative blogging platform in order to ensure that the blogging characteristics will continue to be available. This might include the migration of a live blog to an alternative platform (which would not normally be described as archiving) but could also involve copying the blog's rich content in order to support data mining or other business processes which may not be possible on the original environment.

Production of a physical manifestation of the content: It may be felt desirable to produce a physical manifestation of a blog, such as a hard copy printout, for various purposes, including marketing purposes or to provide access to the content when online access is not possible.

4. TECHNICAL APPROACHS

4.1. HTML Scraping

The HTTrack offline browsing software (see http://www.httrack.com/) has been used to create copies of the UK Web Focus and IWMW 2009 blogs. This approach is simple to use and requires no special access permissions in order to archive public blogs. However the archived resource is a static Web site and the blog's structure (individual blog posts, comments, etc.) is no longer available as a managed resource.

4.2. Blog Migration

An experiment to migrate the rich content of the blog took place in July 2007 [5]. A blog was created on the VOX platform and the content of the blog was migrated

using the host blog's RSS feed. This approach did maintain the structure of the individual posts although comments were lost. However since the migration relied on the host blog's RSS feed this approach is unlikely to be usable for well-established blogs where RSS feeds typically provide access only to recent posts.

An alternative approach is to use the blog service's export functionality and migrate the content to either different blog software or to a platform hosting the same software. This approach has been used to migrate the UK Web Focus blog to another instance of WordPress which demonstrated that not only blog posts and comments could be successfully migrated but also draft posts and embedded objects.

4.3. Processing RSS Feeds

Despite limitations of RSS to provide content for reuse, it is possible on WordPress to provide an RSS feed not just for new posts but also for all views of the blog [3]. This feature is currently being evaluated as a mechanism for migrating blogs if it is not possible to have access to an export file – see [11].

4.4. Production of PDFs

On the second anniversary of the launch of the UK Web Focus blog a PDF version of the blog was created [5]. Although this fails to provide a reusable resource there may be use cases for which this provides an appropriate solution for preserving the content of a blog.

4.5. Physical Manifestation of a Blog

Although the provision of a blog in a physical format (such as a printed book) may appear to be an unusual approach to digital preservation this approach could be of interest for a student or researcher wishing to provide tangible evidence of their blogging output. The Lulu print-on-demand service (see http://www.lulu.com/) is currently being evaluated for the production of hard-copy outputs of our blogs. This will include policy decisions on the content to be published (e.g. should comments be included?).

4.6. Third-Party Web Archiving Services

Commercial Web archiving services such as the UK Web Archive (see http://www.webarchive.org.uk/ukwa/) and Archive-It (see http://www.archive-it.org/) provide an alternative approach to the provision of archives.

The UK Web Archive states that "*If you are the owner of a UK website you are especially encouraged to nominate your own site: this will make the permissions process as straightforward as possible. However, please note that we reserve the right to decide whether to include a site and that for technical reasons we may not be able to archive all sites.*" [15]. The JISC PoWR blog was submitted to the UK Web Archive service. Archives of the site were gathered in January, April, July and October 2009 and January 2010 but none of the updates to the blog made between January and July 2010.

Archive-It is a subscription service which has "*95+ partners include: state archives, university libraries, federal institutions, state libraries, non government non profits, museums, historians, and independent researchers*" [4]. Examining the Archive-It service in July 2010 revealed that only one resource from the JISC PoWR blog was available in the archive.

5. POLICY AND RELATED ISSUES

5.1. Blog Policies

In addition to the evaluation of various technical approaches for the migration of blog content we have also implemented appropriate policy statements regarding the ownership of the content, access to the content and rights if the blog author leaves the host institution or if there are changes to the terms and conditions or sustainability of the third party service.

The blog polices for the UK Web Focus and Ramblings of a Remote Workers blog state that:

"*A copy of the contents of the blog will be made available to UKOLN if I leave UKOLN. Note that this may not include the full content if there are complications concerning their party content (e.g. guest blog posts, embedded objects, etc.), technical difficulties in exporting data, etc.)"* and "*Since the blog reflects personal views I reserve the rights to continue providing the blog if I leave UKOLN. If this happens I will remove any UKOLN branding from the blog*" [10].

5.2. Risk of Use of Third Party Services

A risk assessment approach to use of third party services to support UKOLN activities was first used at the IWMW 2006 event when a risk assessment statement was published which provided an assessment of risks and plans for mitigating against such risks [12]. Risk statements have been produced for subsequent events which ensure that the organisers consider the risks they may be taking and also provides documentation on the third party services which are used.

A framework for assessing the risks of use of third party services has been published which builds on these initial approaches [7].

5.3. Privacy Issues

Possible concerns regarding the publication of photographs of participants at the IWMW 2009 event were identified prior to the event. The event booking form used an approach taken for bookings at recent JISC conferences which stated that photographs would be taken at events. However the event organisers would use their discretion when reusing such photographs. In addition we provided 'quiet area' at the event which was intended for participants who did not wish to be photographed or distracted by the noise of use of laptops [13]. We sought to ensure that photographs used on the blog would not be likely to cause embarrassment. We have also agreed that we will be prepared to remove

photographs from services under our control if a rights-holder expresses their concerns if this can readily be achieved.

5.4. Ownership Issues

In order to clarify ownership issues we use Creative Commons licences for our blogs. The UK Web Focus blog contains the following statement:

"*This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 2.0 UK: England & Wales License. This licence applies to textual content published by the author and (unless stated otherwise) guest bloggers. Comments posted on this blog will also be deemed to have been published with this licence. Please note though, that images and other resources embedded in the blog may not be covered by such licences.*"

Note that the statement acknowledges the complexities of copyright issues. A risk assessment approach is taken based on ideas described in [8].

6. ARCHIVING APPROACHES FOR THE CASE STUDIES

6.1. The JISC PoWR Blog

Our original intention with the JISC POWR blog was to continue to publish occasional posts related to Web preservation issues but at a significantly lower level. Our aim was to allow the blog to be reused if additional funding became available to continue our work in providing advice on best practices for the preservation of Web resources. However although we were successful in obtaining additional funding this covered a broader area that Web preservation. We therefore felt that it was inappropriate to change the scope of the original blog and have chosen to archive the blog.

The process of freezing the JISC PoWR blog involved carrying out an auditing of use of the blog, with a post published containing a summary of the numbers of posts and comments published, numbers of contributing authors, etc.

An audit of the blog technologies used was also carried out and published. This included details of the WordPress plugins installed and theme and widgets used. We became aware of the value of such audits when, in May 2010 the hosting agency upgraded the platform from WordPress 1 to WordPress 2. A consequence of the update was the loss of the theme, with the look-and-feel reverting to the WordPress default. We need to know which theme we had been using in order to recreate the previous appearance of the blog.

In order to have a better understanding of how the blog was used we created a copy of the blog on the UKOLN Intranet. This will enable us to analyse the contents of the blog using a variety of WordPress plugins which are not available on WordPress.com.

The availability of the backup copy of the blog meant that we could change configuration options which we would not want to do on the live blog. We set the number of RSS items provided to a large number so that the entire contents of the blog posts and comments could be made available via an RSS feed. The RSS feed was used to produce a Wordle word cloud which provides a visualization of the contents of the blog and the comments which have been provided. The RSS feed was also processed by Yahoo Pipes. This enabled the contents of the blog to be processed by an RSS to PDF tool, with a series of PDF files being produced in chronological order (with the capability of applying additional filtering if so desired).

A blog post announcing the "*Cessation of posts to the JISC PoWR blog*" was published in July 2010 which helped to ensure that the status of the blog had been provided to visitors to the blog [2].

The archiving approaches taken to the JISC PoWR is summarised as:

A record of the status of a project blog was taken and published. A rich copy of the contents of the blog was held on a Wordpress blog on the UKOLN Intranet which provides a backup managed within the organisation.

6.2. The UK Web Focus Blog

Periodic copies using a rich XML export of the content of the UK Web Focus blog have been created and used to recreate the blog on a Wordpress installation on UKOLN's Intranet.

The ability to configure the backup blog enables additional management and auditing approaches to be carried out on the blog which cannot be implemented on the live blog due to the limitations provided on the WordPress.com or to avoid changing the environment for users of the live blog.

The appearance of the blog has been changed so that all posts are displayed on a single (very large) HTML page. The contents of this page has been copied and pasted into an MS Word file and an automatic table of contents has been generated. The blog can then be managed in a similar fashion to conventional MS Word documents.

The numbers of RSS items which can be accessed had been changed on the backup blog to a large number, to enable all posts to be processed using RSS (on the live blog only the most recent 25 items are available by the blog's RSS feed). Yahoo Pipes can be used to process the complete contents of the blog, with the ability to provide a variety of filtering mechanisms. This approach has been used to provided PDF copies of the blog on an annual basis, using the RSS2PDF (http://rss2pdf.com/) service. The selected view of the blog can then be managed in a similar fashion to conventional PDF documents.

In addition to these in-house approaches the blog was also submitted to the UK Web Archive service. However no notification has been received from the

service and the blog does not appear to have been retrieved by the service.

The archiving approaches taken to the UK Web Focus is summarised as:

Periodic rich copies of the UK Web Focus blog are taken and installed on the UKOLN Intranet for use in more detailed analyses of the blog. The backup can also be used to avoid loss of the content in cases of a lack of sustainability to the master copy.

6.3. The IWMW 2009 Blog

The IWMW 2009 event blog was used in the run-up to the event, during the event and shortly after the event had finished when a number of posts were published after the event summarising the feedback received.

In order to provide clear termination of the blog a post was published which announced its closure [1] in line with advice on best practices for closing blogs published [14].

However since IWMW is an annual event we recognised that we may wish to publish occasional posts linking to the forthcoming event. Since the blog can provide marketing benefits, with links likely to help enhance Google ranking it has been decided that the blog will continue to be hosted on WordPress.com, though with some minor changes:

- A sidebar widget ensures that the status of the blog is clear.
- A widget provides links to key resources related to the event.
- Widgets providing access to dynamic content, such as live Twitter feeds, have been removed.

In the preparation work for the archiving the blog we observed that a number of posts contained embedded objects (such as video clicks hosted on the Vimeo.com service) which did not include a link to the object on the remote service. Since we realized that loss of the embedding mechanism (which is a configurable option in WordPress) would result in loss of the embedded object and no information being provided on the location of the hosted video clips we edited the posts to included a link to the object on the external service as illustrated in Figure 2 (taken from http://iwmw2009.wordpress.com/2009/08/07/take-aways/):

> Note that these three videos are hosted on Vimeo and can be accessed directly at:
> * http://vimeo.com/5976384
> * http://vimeo.com/5976404
> * http://vimeo.com/5976472

Figure 2: Links to embedded objects

We have decided not to keep an XML archive of the blog content since we feel the risk of loss of the content is small and there will be no serious consequences if the content is lost. However we have used WinHTTrack to keep a static copy of the blog which is stored on the UKOLN Intranet.

We have also published a static page on the blog which summarises these policies (see http://iwmw2009.wordpress.com/status-of-this-blog/).

Figure 3: Closure of the IWMW 2009 Blog

An illustration of the home page is shown in Figure 3 with the key features highlighted.

The archiving approaches taken to the IWMW 2009 blog is summarised as:

A static copy of the IWMW 2009 blog is available on the UKOLN Intranet. The backup can also be used in case of a lack of sustainability to the master copy.

7. IDENTIFICATION OF GOOD PRACTICE

The work in understanding appropriate solutions for our archiving of professional blogs hosted in the Cloud has helped us to identify appropriate practices which may be particularly relevant for funding bodies who wish to ensure that project-funded activities which make use of blogs provided by third parties implement appropriate approaches for ensuring that the content provided on such blogs does not disappear unexpectedly.

The checklist we have developed includes the following steps:

Planning: Preparation for archiving blogs should begin before the blog is launched. A blog policy can held to clarify the purpose of the blog and its intended audience.

Clarification of rights: A copyright statement covering blog posts and comments can also minimise the legal risks in archiving the blog.

Monitoring of technologies used: Information on the technologies used to provide the blog, including blog plugins, configuration options, themes, etc. can be useful if a blog environment has to be recreated.

Auditing: Providing an audit of the size of the blog, numbers of comments, usage of the blog, etc. may be useful in helping to identify the value of a blog and in ensuring that interested parties are aware of how well-used the blog was.

Understanding of costs and benefits: The audit should help to inform the decision-making processes regarding the effort which needs to be taken for the selected blog archiving strategy.

Identification and implementation of archiving strategy: The appropriate blog archiving strategy needs to be selected. As illustrated in the case studies this could include 'freezing' a blog on the external service, with an organisational backup copy (in a variety of formats) or the continuation of an active blog, with a backup copy of taken in case of unexpected data loss.

Dissemination: It will be desirable to ensure that end users are aware of the existence of an archived copy. Ideally such information will be made publicly available. The summaries of the approaches taken in the three case studies illustrate that such dissemination work need not be time-consuming to implement.

Learning: During the planning, auditing, selection and implementation of appropriate archiving strategies there are likely to be lessons learnt (such as, in the case of the IWMW 2009 case study the need to include links to external services and not just embed the objects). Such experiences should be used to inform subsequent blogging practices.

Organisational Audit: There is a likely to be a need to carry out an organisation audit of use of blogs held on third party services which may be at risk. Such an audit should initially identify (a) location of such blogs; (b) their purpose(s); (c) the owner(s) and (d) their perceived importance. This information should help to inform decisions on the archiving strategies, along the lines described in this paper.

8. CONCLUSIONS

This paper has reviewed the approaches which have been taken to facilitating long-term access to blogs hosted in the Cloud which are used to support professional activities.

The need to ensure that preservation policies are developed and implemented by JISC-funded projects has been described in [9]. Since many of the blogs provided by JISC-funded development projects may be hosted on third-party services there is a need to document and share the variety of possible technical approaches to the migration of content and related policy issues.

The approaches which have been described seek to address the difficulties which organisations are likely to experience in adopting similar approaches, including the potential difficulties of motivating content providers of the need to address such preservation issues and the limited resources which is likely to be available to implement such practices.

9. REFERENCES

[1] Guy, M. *Last Orders at the IWMW2009 blog*, IWMW 2009 blog, 12 August 2009, <http://iwmw2009.wordpress.com/2009/08/12/last-orders-at-the-iwmw2009-blog/> 2009

[2] Guy, M. *Cessation of posts to the JISC PoWR blog*, JISC PoWR blog, 19 July 2010, <http://jiscpowr.jiscinvolve.org/wp/2010/07/19/cessation-of-posts-to-the-jisc-powr-blog/> 2010

[3] Hirst, A. *Single Item RSS Feeds on WordPress blogs: RSS For the Content of This Page*, OUseful blog, 8 July 2009, <http://blog.ouseful.info/2009/07/08/single-item-rss-feeds-on-wordpress-blogs-rss-for-the-content-of-this-page/> 2010

[4] Internet Archive, *Archive-It*, <http://www.archive-it.org/public/faq.html> 2010

[5] Kelly, B. *A Backup Copy Of This Blog*, UK Web Focus blog, 19 April 2007, <http://ukwebfocus.wordpress.com/2007/07/19/a-backup-copy-of-this-blog/> 2009

[6] Kelly, B. *The Second Anniversary of the UK Web Focus Blog*, UK Web Focus blog, 31 October 2008, <http://ukwebfocus.wordpress.com/2008/10/31/the-second-anniversary-of-the-uk-web-focus-blog/> 2008

[7] Kelly, B., Bevan, P., Akerman, R., Alcock, J. and Fraser, J. *Library 2.0: Balancing the Risks and Benefits to Maximise the Dividends*, Program (2009), Vol. 43, No. 3, pp. 331-327. <http://opus.bath.ac.uk/15260/> 2009

[8] Kelly, B. and Oppenheim, C. *Empowering Users and Institutions: A Risks and Opportunities Framework for Exploiting the Social Web*, CULTURAL HERITAGE online conference, Florence, 15-16th December 2009. <http://opus.bath.ac.uk/17484/> 2009

[9] Kelly, B. *The Project Blog When The Project Is Over*, Web Focus blog, 15 March 2010, <http://ukwebfocus.wordpress.com/2010/03/15/the-project-blog-when-the-project-is-over/> 2010

[10] Kelly, B. *Blog Policies*, UK Web Focus blog, <http://ukwebfocus.wordpress.com/blog-policies/> 2010

[11] Slideshare. *UK Web Focus Blog Posts 2009*, <http://www.slideshare.net/lisbk/uk-web-focus-blog-posts-2009> 2009

[12] UKOLN. *Risk Assessment For The IWMW 2006 Web Site*, <http://www.ukoln.ac.uk/web-focus/events/workshops/webmaster-2006/risk-assessment/> 2006

[13] UKOLN. *Quiet Area, IWMW 2009*, <http://iwmw.ukoln.ac.uk/iwmw2009/quiet/> 2009

[14] UKOLN, *Closing Down Blogs*, Cultural Heritage briefing document no. 81, March 2010, <http://www.ukoln.ac.uk/cultural-heritage/documents/briefing-81/>, 2010

[15] UK Web Archive, *FAQ*, <http://www.archive-it.org/public/faq.html#605> 2010